Stanford University Libraries and Academic Information Resources
Preservation Assessment of Digital Objects

Part 1: The TAG Team Questionnaire

Background
In the early stages of developing a digital preservation program at Stanford, it was clear that the Stanford Digital Repository (SDR) would need to offer a range of services to its prospective clients. While the initial streams of content to be preserved were known to be highly normative and predictable, indications of the varied nature of the materials, limited resources, and enormous volume of files to be preserved in the future necessitated a tiered approach to repository services, such as metadata encoding, pre-ingestion transformations, long-term format migration and delivery, in addition to bit preservation.

A team of individuals -- the Technical Assessment Group, or TAG Team -- came together to develop a framework for categorizing digital objects to be preserved. The thinking was that such a framework would not only prevent the SDR from becoming an undifferentiated heap of content, but that it would also further the development of administrative principles and policies around which the SDR infrastructure and service model would grow.

In approaching the task of developing tiered services for categories of digital objects, the following questions immediately arose:

- How will inevitable change affect the nature of the digital objects stored in the SDR? What may become lost in the process of migrations and transformations?
- What attributes, if any, of an object are crucial to its on-going use and value as an information resource?
- What are the underlying technical characteristics of an object that may prevent those attributes from being preserved?
- What external (non-technical) factors, if any, may have a bearing on the extent of services appropriate for a digital object or collection?

A mechanism was needed to raise questions and to record the answers, a tool to gather vital information about an object, or groups of like objects, and to explicitly express the intent or will of the “content owner”, the person who best knows the information resource, its creation, and its value. The tool needs to accomplish these functions within the context of what is more or less known to be technically possible with respect to digital preservation, and then assess the interplay of these factors in order to set reasonable expectations for both content owners and repository managers about the prognosis for maintaining accessibility to the information encoded in the digital objects over time.

The Questionnaire
The group spent several months identifying the key issues, defining terms, reviewing the literature[1], studying formats and metadata elements, and consulting with colleagues. With

the results of these efforts, we drafted a questionnaire that incorporated the laundry list of factors exposed in our research and discussions that may impede digital preservation, from a technical perspective, or otherwise may impact long-term management of the objects, from a collection administration perspective. The questionnaire also was designed to serve a secondary role in determining the degree of metadata (primarily administrative, including technical, some structural, and other preservation metadata) that may be necessary to adequately document some objects for long-term management. It focused on assessing the types of normative objects to be initially ingested in the SDR (i.e., TIFFs, ASCII text files, PDFs), but it was expected that the framework that the questionnaire represented could be extended in the future to other object types (e.g., audio, video, web, complex, etc.).

In order to manage the results of the survey, we decided upon the notion of a scale of complexity, a horizontal spectrum on which the relative “preservability” of an object can be gauged based upon its technical risk factors. For each risk factor revealed in or exhibited by an object, one point is scored. The more points scored for a given object, the increasingly complex its preservation is expected to be. The group felt that weighting the scores based on the questions was beyond our immediate technical expertise, and that in fact it might not be practical due to the unpredictability of the technological landscape in which digital preservation activities take place. Answers to questions aimed at exposing external, non-technical factors (such as circumstances of origin and acquisition, retention expectations, uniqueness/rarity, etc.) were not scored at all, because it was felt that additional input from curators as well as experience from further collections case studies was required before these factors could be carefully evaluated for their applicability.

Given that at the time of the questionnaire’s completion, the SDR was still very much in prototype and had no dedicated management, development or production staff, the practical implementation of the questionnaire remained an open question. Under these circumstances, it was conceived that the completion of the questionnaire would be an iterative process mediated by a “repository liaison” (i.e., someone with technical expertise and who is involved in the production end of digital library projects) in consultation with the content owner. The score, if not the entire questioning process itself, would inform the negotiation of a service agreement. A web-based form was created (access limited to Stanford only), with an Oracle table as a backend to store the answers and scores, in order to demonstrate the concept.

The questionnaire is outlined in the attached document, `TAGQuestionnaireAIHT.pdf.` It is organized into the following sections:

Non-Technical Factors:	All formats
Technical Factors:	All formats
	Image
	PDF
	Text

Part II: Toward Automation: The AIHT Project

From the completion of the TAG Team questionnaire in April 2003, its implementation has continued to remain an open question. The questionnaire has served a role as a theoretical framework which is influencing the structure of developing SDR services. As the Stanford team has explained, with the AIHT project we intended to continue developing capabilities for assessing the preservability of information encoded in files preserved in the SDR through extending the concepts the questionnaire represents, and we have achieved that goal.

Shifts in the Approach

In order to apply the questionnaire's concepts to a large, heterogeneous collection like the GMU 9/11 Archive, it was obviously necessary to automate most, if not all, of the file assessment process. The work of two other organizations working in the digital preservation community contributed to Stanford's ability to automate preservation assessment of digital objects.

One break-through that supported our ability to automate was the availability of JHOVE, the tool created by Harvard that provides automated verification and identification, not to mention extraction of technical metadata, of a number of key file formats. With the possibility that JHOVE could expose technical vulnerabilities of specific files automatically, it was possible to abandon the instance-level approach which had been embedded in the questionnaire, in favor of a broader approach where an object is assessed more generally along the lines of its format type. A broader approach was key to automating the assessment of digital files; it follows that a more general framework of the factors that impact the preservability of information within formats was necessary. For this, we turned to the work of the Library of Congress Office of Strategic Initiatives as presented by Carl Fleischauer and Caroline Arms at the Digital Library Federation Fall Forum 2003 (see: http://memory.loc.gov/ammem/techdocs/digform/ and http://www.diglib.org/forums/fall2003/fallforum03.htm#p1).

The presentation titled "Digital Formats: Factors for Sustainability, Functionality, and Quality" describes the results of their work to "provide information to help LC staff develop strategies and practices for incoming [digital] content . . . by identifying preferred formats" (Arms and Fleischauer 2003). Two types of factors, *sustainability* and *quality & functionality* factors, emerged as the primary forces which have a bearing on whether or not a format can be considered preferable to others. Of interest is the table in which the factors making up the anticipated sustainability of a file format – disclosure, adoption, transparency, self-documentation, external dependencies, patents, and technology protection measures – are measured and analyzed against a handful of specific formats. This approach relates to Stanford's work, because it effectively generalizes and categorizes much of the spirit and some of the substance of the TAG Team questionnaire. We adopted it, in large part, and developed a matrix of our own for the analysis of predominant formats. It is this matrix that now serves as the basis underlying SDR preservation assessment activities and developing policies.

The Format Scoring Matrix

The Format Scoring Matrix is contained and described in greater detail in the attached document, `FormatScoringMatrix.pdf`. Having evolved over the course of the AIHT project, the matrix at this stage is in its most formal and developed state. While it appears to serve as a reasonable measure of a format's sustainability based on current knowledge, testing is required and additional revisions are likely. Several anticipated developments include the inevitable need to break out the various types of marked-up text for a more granular analysis of the distinctions between them. A close analysis of datasets is also called for to determine how the matrix can accommodate them. Further research and experimentation is necessary with respect to the final analysis of formats against the set of sustainability factors. Finally, the as-of-yet unexplored impact of relationships between highly complex compound objects will need to be accommodated in the overall SDR preservation assessment process. In the very near future, Stanford will be closely examining formats used for geospatial data, formats inherently more complex than those already addressed in our process. The matrix, indeed all rules which frame file assessment, will always be evolving.

It is worth noting that not all of the factors identified by the Library of Congress were adopted for use in Stanford's matrix. Excluded from the format analysis are two sustainability factors: "impact of patents" and "technology protection mechanisms". As Arms and Fleischauer acknowledge in their discussion, the topic of patent impact is a tricky one and needs further exploration. Even among some "standard" formats, the impact of patents can be felt. For the time being, the SDR assumes that the degree to which a format is encumbered by patents, or a similar formal claim on technological invention or innovation, directly affects the other sustainability factor of external dependencies and perhaps indirectly affects the factor of adoption.

With respect to technology protection mechanisms, they only have a bearing on the extent of services that the SDR can offer if, and only if, a specific file's internal technology protection mechanism is enabled. This state should be revealed and accounted for during routine file analysis processes prior to ingest. Therefore the potential for a file to have internal protections does not serve a purpose in the Format Scoring Matrix.

Also, the *quality and functionality* factors were not explicitly adopted in Stanford's Format Scoring Matrix. Such factors "pertain to current and future usefulness, e.g., for scholarship or repurposing". Specific characteristics indicative of a format's potential for quality and functionality include support for high resolution, color management, document structure and navigation, etc. (Arms and Fleischauer 2003). At Stanford, characteristics of quality and functionality are considered separately in the determination of SDR "preferred" formats (see below).

Part III: The SDR Preservation Assessment Process, version .01

Preservation assessment of files to be stored in and managed by the SDR occurs in several steps in the course of preliminary preparation of files for ingestion, a process carried out by a file-traversing program developed at Stanford called “The Empirical Walker.” The steps can be generally described as follows (refer also to the diagram depicted in WorkflowSDRDigiProv.pdf):

1. Initial File Identification and MIME Type Assignment: The Empirical Walker is run on a specified collection of digital objects. As it traverses the directory(s) of files, it identifies file extensions and maps them to corresponding MIME types.

2. File Validation: A file validator (e.g., JHOVE) is invoked for those files for which there exists an applicable tool or module. Output is stored temporarily for subsequent processing.

3. Preservation Assessment Process Initiated: The Empirical Walker assigns a default Format Status by checking a registry of values, numbers 0-5, derived from the Format Scoring Matrix. Default Format Scores are grouped and matched with corresponding Preservation Quality Status levels. As the name implies, the Preservation Quality Status provides a relative qualitative measure as a result from the format assessment and serves as a useful gauge in subsequent file analysis.

   The complete results of this analysis are stored as preservation metadata in the METS Digital Provenance Section.

4. File Analysis: Analysis has two primary goals. The first goal is to examine the output of the file validation process in search of:
   a. invalid or not well-formed files;
   b. technical characteristics which could pose potential complications in preservation activities (such as a TIFF that is compressed);
   c. technical characteristics which indicate that the file has reduced preservation-risk associated with its default format status (such as a PDF which meets the anticipated PDF-A profile).

   A set of rules guides the Empirical Walker to flag any output pertaining to a specific file that could alter its preservation prospects, which up to this point are based solely on a general assessment of file formats articulated in the File Scoring Matrix.[2]

   The second goal of supplemental analysis is to identify those files that may benefit from transformation or normalization. The Empirical Walker determines if the extension represents a MIME type or format that has particular characteristics at risk of loss in future digital preservation activities and therefore could be transformed pre-ingest by means of reformatting or normalization in order to improve its preservation prospects. For instance, a Photoshop document (*.psd), which has a low format status (4), could be reformatted as a PNG or TIFF with little to no loss, and thus earn a higher preservation assessment as a result. Similarly a MS Word document could be reformatted as plain text for enhanced preservation of the textual content over time.

All file analysis results, including the benefits and risks resulting from any optional transformations, such as changes to the file's content, functionality, or look-and-feel, are reported to repository staff and/or the content owners (see below).

5. Preservation Policy Status Assignment: As a result of the file analysis steps, a suggested policy status level is assigned to the file. This value guides the SDR staff and content owners in the negotiation of an on-going preservation service agreement to be applied to the files or collection of files. There are five policy status levels: preferred, approved, acceptable, minimal, unknown.

   A class of preferred formats (as opposed simply to "approved") is necessary for business reasons because it focuses the number of formats for which the SDR is committed to providing full support services. Not all formats with a Format Score of zero automatically earn the status of a preferred format; the capacity of a format for enhanced quality and functionality must be factored in. Similarly, a Format Score of zero is not required to earn the "preferred" status; a format that is both highly suited to a specific purpose within the digital library context *and* free of risk factors does not always exist.

SDR Preferred Formats

Plain text	ASCII
	UTF-8
Marked-up text	XML 1.0
Image	TIFF 5.0 and above (uncompressed)
Page-Viewer	PDF* (any version)
Audio	WAVE (linear pulse code modulation)
Video	*TBD*

*Despite the PDF format's lack of transparency and external dependencies, factors that give it a "medium", not high, Preservation Quality score, PDF is currently Stanford's preferred page viewing format, because it is the *de facto* standard with extremely wide market penetration.

For those files that do not qualify as preferred, another status level is assigned according to the Preservation Quality Status of the file. The following table outlines the correlation between scores and statuses.

Default Format Score	**Preservation Quality Status**	**Policy Status**
0-1	High	Approved
2-3	Medium	Acceptable
4	Low	Minimal
5	Low	Unknown

Written policies and specific services associated with the various policy levels remain under development.

6. Reporting: At the culmination of the file analysis, a report is issued to repository staff and/or content owners/depositors. The report includes the results and scores of the various analytical tests to which the files have been subjected. Those files with traits, exposed in supplemental analysis, which may require a modification of policy status or other subsequent action are indicated in the report. Possible outcomes reflected in the report include:
    a. *Change in status*: a file has been demoted or promoted from its default Preservation Quality status, influencing the type and extent of preservation services it qualifies for;
    b. *Transformation options*: possible target formats for an original file format are outlined, detailing any potential loss or gain associated with such action with respect to a file's contents, formatting, functionality, etc.;
    c. *Red flags*: a file has a particular technical trait or quality that is noteworthy but requires additional human input to determine an appropriate action, if any.

Preliminary Conclusions

While several parts of this process remain to be built and tested in full, we believe this multi-stage preservation assessment process provides an advantage of flexibility due to its modular design. Twenty-one of forty-one possible technical risk factors from the questionnaire have been identified or exposed in an automated fashion by way of this assessment process (they are indicated as such in `TAGQuestionnaireAIHT.pdf`). With the expanded capability of file validation tools in the future, it is expected that this number will increase over time. Through the incorporation of reporting to content owners results from the analysis steps, it will be possible to seek input from the repository client about the potential presence of specific characteristics of documents (such as links or other embedded interactivity and multimedia in proprietary file formats) that are at risk of loss over time but are not easily identified through automation, to manage the client's expectations for long-term preservation of content more generally, and to inform the client about transformation options or other repository services that may be appropriate for the files in question. The goal is for the reporting aspect to create a key role for the content owner by shifting some of the weight that the preservation assessment and decision-making process entails from SDR staff to the content owner. This goal is in keeping with our overall strategy to automate as much of the pre-ingestion file preparation process as possible and to push service-oriented tools to the repository clients wherever possible.

[1] Key sources consulted which informed the overall development of the questionnaire include: Bennett, John C. *JISC/NPO Studies on the Preservation of Electronic Materials: A Framework of Data Types and Formats, and Issues Affecting the Long Term Preservation of Digital Material*, British Library Research and Innovation Report No. 50, 1997, <http://www.ukoln.ac.uk/services/elib/papers/supporting/pdf/rept011.pdf>; Harvard Office for Information Systems *DRS Policy Guide*, 5 October 2001 <http://hul.harvard.edu/ois/systems/drs/policyguide.html>; and LeFurgy, William G. "Levels of Service for Digital Repositories," *D-Lib Magazine*, May 2002, Volume 8 Number 5, <http://www.dlib.org/dlib/may02/lefurgy/05lefurgy.html>.

[2] Evaluation of errors in HTML documents revealed by JHOVE is still on-going. Until complete testing is carried out, the file assessment process simply flags those HTML documents that are invalid or not well-formed; details or recommended outcomes have not yet been incorporated into the supplemental assessment and reporting process.